

Liquid Foundation Models: Our First Series of Generative AI Models

Published September 30th, 2024



Announcing the first series of Liquid Foundation Models (LFMs) – a new generation of generative AI models that achieve state-of-the-art performance at every scale, while maintaining a smaller memory footprint and more efficient inference.

Try Liquid

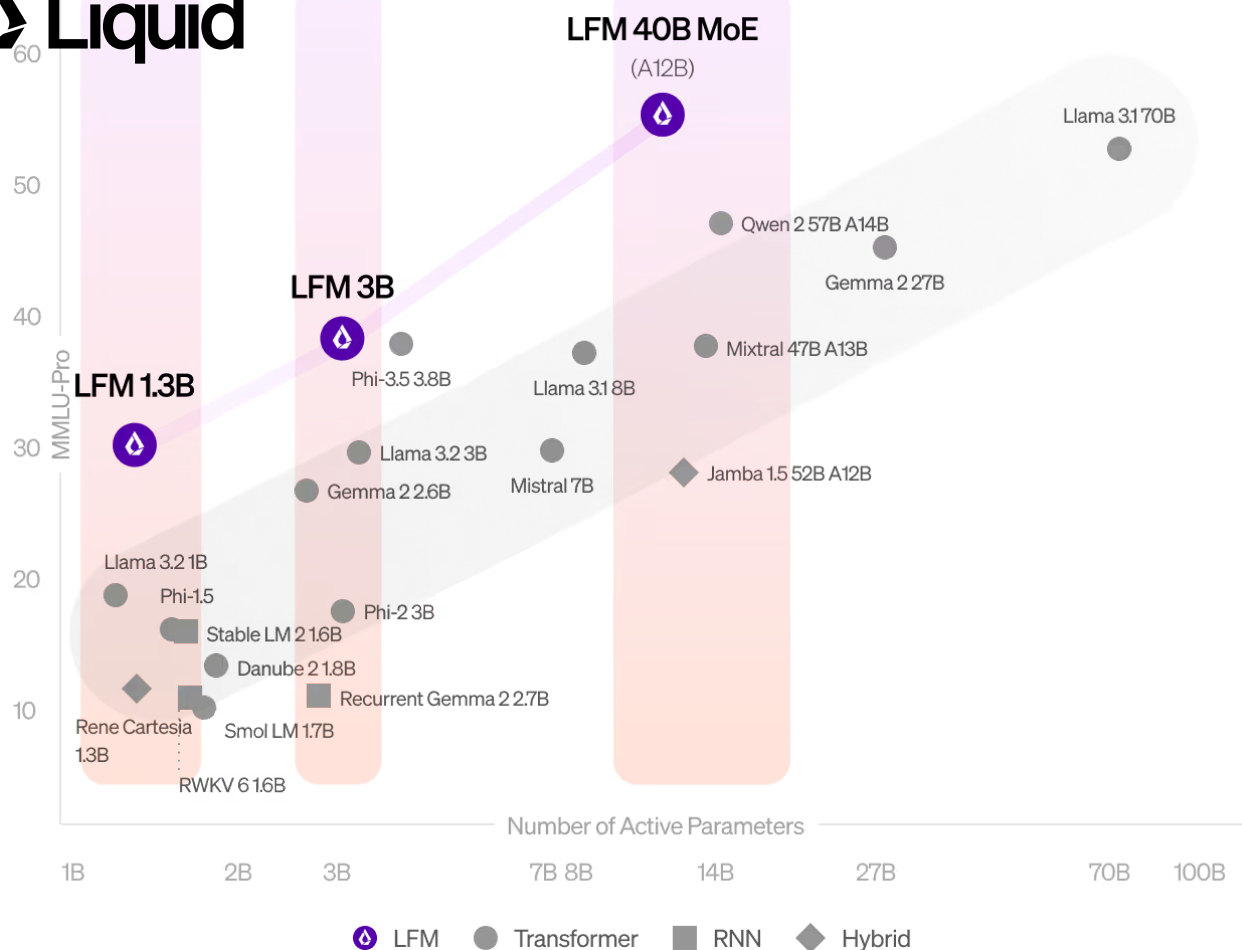


Fig. 1. LFM models offer a new best performance/size tradeoff in the 1B, 3B, and 12B (active parameters) categories.

Takeaways

Introducing the First
Generation of Language
LFMs

[Reimagining Model
Architectures](#)


Join us as an early
adopter of LFM models

Takeaways

We announce the first series of Liquid Foundation Models (LFMs), a new generation of generative AI models built from first principles.

Our 1B, 3B, and 40B LFM models achieve state-of-the-art performance in terms of quality at each scale, while maintaining a smaller memory footprint and more efficient inference.



Try LFM today on **Liquid Playground**, **Lambda (Chat UI and API)**, **Perplexity Labs**, and soon on **Cerebras Inference**. The LFM stack is being optimized for NVIDIA, AMD, Qualcomm, Cerebras, and Apple hardware. 

We build private, edge, and on-premise AI solutions for enterprises of any size.

We are scaling LFM and expect to introduce new and better capabilities across various industries, such as financial services, biotechnology, and consumer electronics.

[Try Liquid](#)

At Liquid AI, we build new methods for designing powerful AI systems over which we have significant control. We design them the same way engineers built engines, cars, and airplanes: from first principles. Our mission is to create best-in-class, intelligent, and efficient systems at every scale – systems designed to process large amounts of sequential multimodal data, to enable advanced reasoning, and to achieve reliable decision-making.

Today, we introduce the first generation of Liquid Foundation Models (LFMs). LFMs are large neural networks built with computational units deeply rooted in the theory of dynamical systems, signal processing, and numerical linear algebra. This unique blend allows us to leverage decades of theoretical advances in these fields in our quest to enable intelligence at every scale. LFMs are general-purpose AI models that can be used to model any kind of sequential data, including video, audio, text, time series, and signals.

Introducing the First Generation of Language LFM_s

We are proud to release our first series of language models:

A dense **1.3B model**, ideal for highly resource-constrained environments.

A dense **3.1B model**, optimized for edge deployment.

A **40.3B Mixture of Experts (MoE) model**, designed for tackling more complex tasks.

Architecture work cannot happen in a vacuum – our goal is to develop useful models that are competitive with the current best-in-class LLMs. In doing so, we hope to show that model performance isn’t just about scale – it’s also about innovation.

State-of-the-Art Performance

We report the results of our fine-tuned LFM_s and compare them with similar-sized language models using Eleuther AI’s lm-evaluation-harness v0.4. Unless specified otherwise, we compare to other fine-tuned models.



LFM-1B achieves the highest scores across various benchmarks in the 1B category, making it the new state-of-the-art model at this size. This is the first time a non-GPT architecture significantly outperforms transformer-based models.

Benchmark	LFM-1B Preview 1.3B	OpenELM (Apple) 1.1B	Llama 3.2 (Meta) 1.2B	Phi 1.5 (Microsoft) 1.4B
Context length (tokens)	32k	1k	128k	2k
MMLU (5 shot)	<u>58.55</u>	25.65	45.46	42.26
MMLU-Pro (5 shot)	<u>30.65</u>	11.19	19.41	16.80
Hellaswag (10-shot)	67.28	71.8	59.72	64.03
ARC-C (25-shot)	<u>54.95</u>	41.64	41.3	53.75
GSM8K (5-shot)	<u>55.34</u>	0.38	33.36	31.61

LFM-3B delivers incredible performance for its size. It positions itself as first place among 3B parameter transformers, hybrids, and RNN models, but also outperforms the previous generation of 7B and 13B models. It is also on par with Phi-3.5-mini on multiple benchmarks, while being 18.4% smaller. LFM-3B is the ideal choice for mobile and other edge text-based applications.

Benchmark	LFM-3B Preview 3.1B	Gemma 2 (Google) 2.6B	Zamba 2 (Zyphra) 2.7B	AFM Edge (Apple) 3B
Context length (tokens)	32k	8k	-	32k
MMLU (5 shot)	66.16	56.96	56*	60.64*
MMLU-Pro (5 shot)	<u>38.41</u>	27.32	-	-
Hellaswag (10-shot)	78.48	71.31	76*	55.24*



ARC-C (25-shot)	63.99	57.94	56*	45.39*
GSM8K (5-shot)	70.28	44.28	-	-

*Scores reported by the developers. All the other scores were calculated with the same evaluation harness we used for our own models.

LFM-40B offers a new balance between model size and output quality. It leverages 12B activated parameters at use. Its performance is comparable to models larger than itself, while its MoE architecture enables higher throughput and deployment on more cost-effective hardware.

Benchmark	LFM-40 Preview 40B A12B	Jamba 1.5 (AI21) 52B A12B	Mixtral (Mistral) 47B A13B	Qwen 2 (Alibaba) 57B A14B
Context length (tokens)	32k	256k	8k	32k
MMLU (5 shot)	78.76	59.57	73.42	75.75
MMLU-Pro (5 shot)	<u>55.63</u>	28.69	38.12	47.47
Hellaswag (10-shot)	82.07	77.16	<u>87.54</u>	85.96
ARC-C (25-shot)	67.24	60.90	71.33	66.89
GSM8K (5-shot)	76.04	46.47	64.22	77.79

*Scores reported by the developers. All the other scores were calculated with the same evaluation harness we used for our own models.

LFMs are Memory-Efficient

LFMs have a reduced memory footprint compared to transformer architectures. This is particularly true for long inputs, where the KV cache in transformer-based LLMs grows


linearly with sequence length. By efficiently compressing inputs, LFM_s can process longer sequences on the same hardware. For example, compared to other 3B-class models, LFM_s maintain a minimal memory footprint. 

Fig. 2. Total inference memory footprint of different language models vs. the input+generation length.

LFMs Truly Exploit their Context Length

In this preview release, we have optimized our models to deliver a best-in-class 32k token context length, pushing the boundaries of efficiency for our size. This was confirmed by the RULER benchmark, where a length is considered “effective” when its corresponding score is higher than 85.6 [Hsieh et al. 2024 - RULER]. The following table compares several models at different context lengths.

Model	Claimed length	Effective length	4k	8k
Gemma 2 2B (Google)	8k	4k	<u>88.5</u>	0.60
Llama 3.2 3B (Meta)	128k	4k	<u>88.7</u>	82.4
Phi-3.5 3.8 B (Microsoft)	128k	32k	<u>94.3</u>	<u>91.7</u>
Llama 3.1 8B (Meta)	128k	32k	<u>95.5</u>	<u>93.8</u>
LFM-3B	32k	32k	<u>94.4</u>	<u>93.5</u>

This highly efficient context window enables long-context tasks on edge devices for the first time. For developers, it unlocks new applications, including document analysis and summarization, more meaningful interactions with context-

Our goal is to keep scaling LFM's across model size, train/test time compute, and context length. Beyond our language LFM's, we have designed models for various data modalities, domains, and applications that we plan to release in the next months.

Advancing the Pareto Frontier of Large AI Models

To achieve these results, we optimized our pre- and post-training pipelines and infrastructure to ensure our models excel across five criteria:

Knowledge capacity



Multi-step reasoning



Long context recall



Inference efficiency



Training efficiency



Building on a long line of research in designing expressive and efficient learning systems, we have developed a new design space for foundation models, focusing on different modalities and hardware requirements. Our goal is to explore ways to build foundation models beyond Generative Pre-trained Transformers (GPTs).

With LFMs, we put into practice new principles and methods guiding model design, developed by our team over the past months.

LFMs are composed of structured operators. >

LFM architectures are under control. >

LFMs are adaptive and can serve as the substrate for AI at every scale. >

Fig. 3. Our architectures feature custom computational units arranged in *depth groups* (targeted weight sharing), with additional *featurizer interconnections* (feature sharing).

Liquid's design space is primarily defined by featurization and footprint of architectures and their core operators. Featurization refers to the process of converting input data (e.g., text, audio, images, video) into a structured set of

features or vectors that are used to modulate computation inside the model in an adaptive manner. For example, audio and time series data generally requires *less* featurization in operators due to lower information density, compared to language and multi-modal data. The other key dimension is the computational complexity of the operators. Being able to traverse and complete the design space of structured adaptive operators allows us maximize performance with controlled computational requirements.

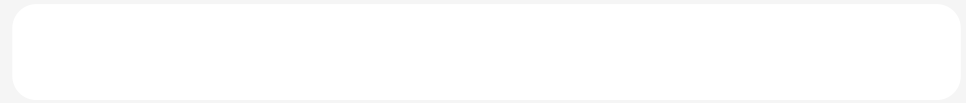


Fig. 4. We built the foundations of a new design space for computational units, enabling customization to different modalities and hardware requirements.

At their core, LFMs are built with computational units that can be expressed as adaptive linear operators whose actions are determined by inputs. The LFM design framework unifies and subsumes a wide range of existing computational units in deep learning, providing a systematic approach to exploring the space of architectures. Specifically, our analysis informs model building by improving three key aspects: token-mixing structure (how the operator mixes embeddings in the input sequence), channel-mixing structure (how it mixes channel dimensions), and featurization, responsible for modulating computation based on the input context.

Join us as an early adopter of LFMs

As we are still in the early stages of this journey, we welcome the opportunity to collaborate and discover the strengths and weaknesses of these systems together.

What are Language LFMs good at today:

- General and expert knowledge
- Mathematics and logical reasoning
- Efficient and effective long-context tasks
- Their primary language is English, with secondary multilingual capabilities in Spanish, French, German, Chinese, Arabic, Japanese, and Korean

What are Language LFM's not good at today:

- Zero-shot code tasks
- Precise numerical calculations
- Time-sensitive information
- Counting r's in the word "Strawberry"!
- Human preference optimization techniques have not been applied extensively to our models yet.

At Liquid AI, we take an open-science approach. We have and will continue to contribute to the advancement of the AI field by openly publishing our findings and methods through scientific and technical reports. As part of this commitment, we will release relevant data and models produced by our research efforts to the wider AI community. We have dedicated a lot of time and resources to developing these architectures, so we're not open-sourcing our models at the moment. This allows us to continue building on our progress and maintain our edge in the competitive AI landscape.

If your enterprise is looking to experience the forefront of AI, we invite you to [get in touch with us](#). If this aligns with your personal goals and ambitions, we invite you to [join our team](#) and drive this vision forward. We are very early on this journey and actively innovating across various aspects of foundation model development and deployment. We invite enthusiastic users to share their experience as well as criticism, and join our red-teaming efforts to improve the capabilities of our models.

[Share your feedback](#)

OCTOBER 23, 2024 | CAMBRIDGE, MA

Come join us at MIT Kresge, Cambridge, MA on October 23rd 2024, to learn more about Liquid as we unveil more products and progress on LFMs and their applications in consumer electronics, finance, healthcare, biotechnology, and more!

[RSVP Here](#)

About Liquid



Liquid AI, an MIT spin-off, is a foundation model company headquartered in Boston, Massachusetts. Our mission is to build capable and efficient general-purpose AI systems at every scale.

Share:





[Press inquiry](#)

[Media kit](#)

[Privacy Policy](#) | [Terms & Conditions](#)

314 Main St, Cambridge, MA 02142

© 2024, Liquid AI, Inc. All rights reserved.